

Kerstin Kleese van Dam

James Carson, David Cowley, Daniel Einstein, Ian Gorton, Andrew Kuprat, Dongsheng Li, Guang Lin, Yan Liu, Lori Ross O'Neil, Jian Yin, Lou Terminello, Suntharampillai Thevuthasan, - PNNL

Shoaib Sufi, Glen Drinkwater, Louisa Casely-Hayford, Brian Matthews - STFC

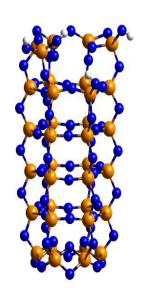
October 2010



Proudly Operated by Battelle Since 1965

Motivation

- Enabling improved analysis through more complete information, facilitated by allowing users to get rapid access to current and past data, related projects, publications etc.
- Enabling rational design and synthesis of new chemical, biological, and materials systems through integrated molecular scale imaging technologies, real time analysis and manipulation.



Capturing Data and Making it Accessible – STFC 2002-2008





UK Science and Technology Facilities Council (STFC)

STFC (formerly CCLRC) facilitates the access to large scale experimental and computational facilities for the UK research community, both through subscriptions to international institutions and by operating a range of world class facilities e.g.:

ISIS Neutron and Muon Facility

Daresbury Laboratory Synchrotron

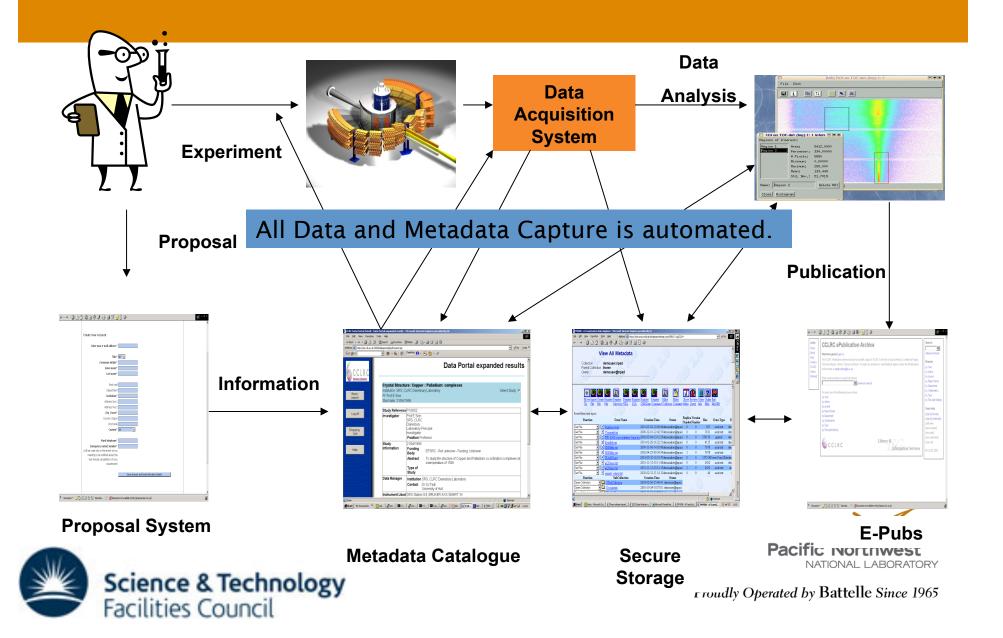
DIAMOND Light Source

Central Laser Facilities

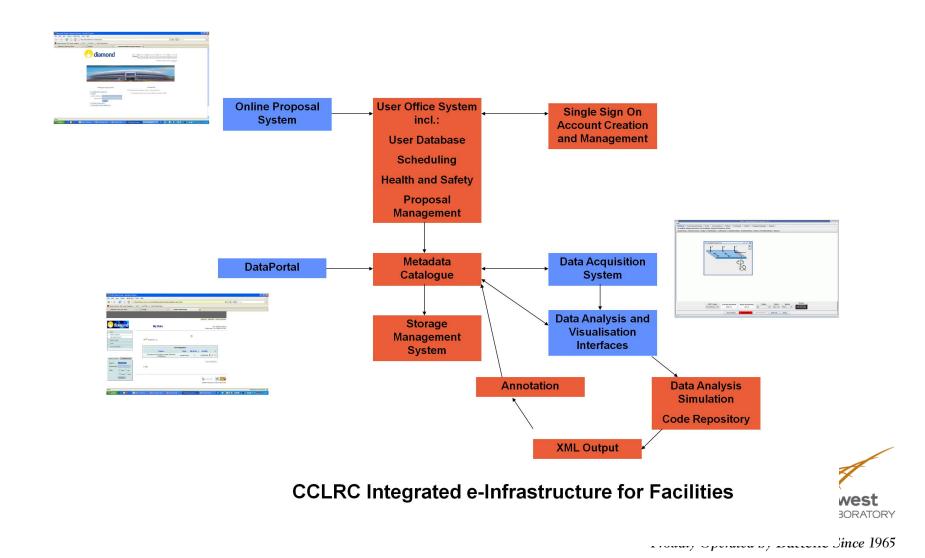




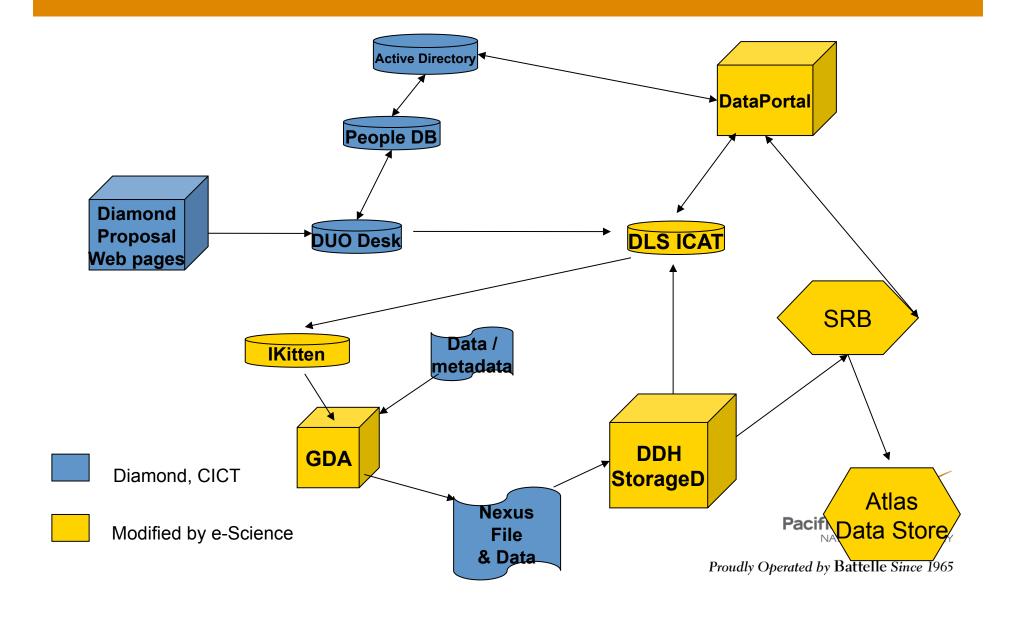
Integrated e-Infrastructure Vision



Integrated Infrastructure Architecture

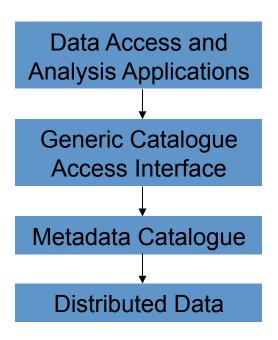


DIAMOND Implementation



ICAT Software Suite

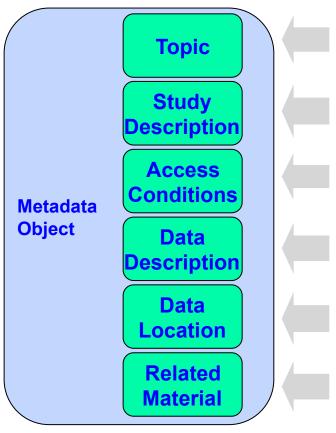
- The ICAT software suite centrally catalogues all experiment related information and extracts key results.
- Where ever possible information is gathered automatically trough integration with existing IT systems such as proposal systems or data acquisition.
- The catalogue and the data it references are accessible via a well defined API for easy embedding into any applications.







Metadata Model



Keywords providing a index on what the study is about.

Provenance about what the study is, who did it and when.

Conditions of use providing information on who and how the data can be accessed.

Detailed description of the organisation of the data into datasets and files.

Locations providing a navigational to where the data on the study can be found.

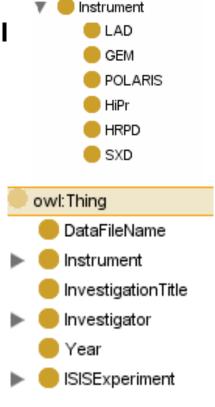
References into the literature and community providing context about the study.





Ontology Support

- 1,700,000 distinctive keywords ISIS ICAT
- These keywords are used to index experimental studies
- The creation of ontology's at ISIS aids the mapping of familiar terms in one domain as well as related concepts in different domains.
- Facilitates searching of data by category and grouping of data into keywords across studies.
 Faster results and enabling of cross facility search.





Pacific Northwest

Infrastructure – Access to Multiple Facilities







CSL - Canada



Data Portal



















What was achieved by 2008

- Agreed common metadata model, data formats and single sign on allows scientists to have rapid access to their work
- A 20 year back catalogue of ISIS raw data
- All future data collected at STFC Facilities and DLS will be curated and made available for reuse now and in the future
- Raw Data archived and available to project collaborators on- and off site within minutes of collection





Lessons Learned

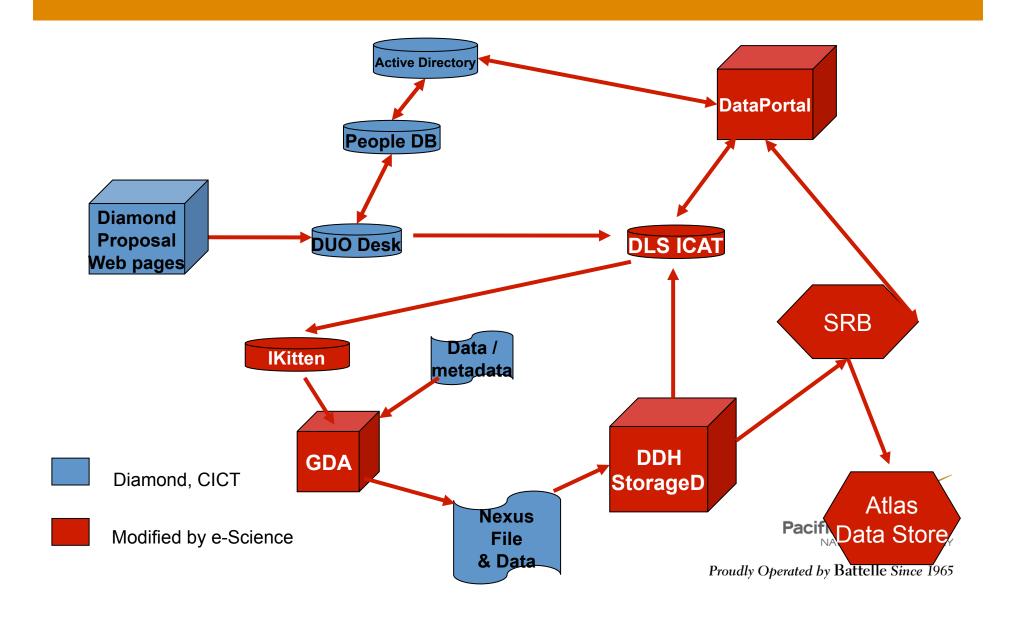
Address Early on:

- Resolve institutional and/or personal issues in dealing with mistrust and the use of the data collected
- Developing agreements, policies and procedures for data governance, ownership, sharing, citations and authorship. How do you reference my data in your publication, agreed embargo times
- Acknowledge 'language barriers' between different project partners

Design:

- Metadata is the key enabling technology
- Automation and reliability of processes are vital
- Interfaces should be as familiar as possible
- Close integration into existing scientific processes
- Step wise progression to take users 'along'

Many Components to Monitor and Control



Automation and Integrated Analysis – MyEMSL 2010





EMSL

Environmental Molecular Sciences Laboratory:
A national user facility integrating experimental and computational resources for discovery and technological innovation



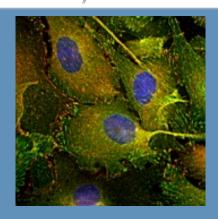




The user program is focused on three science themes

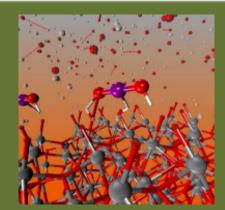


Biological Interactions and Dynamics



Understanding and optimizing the response of biological systems to their environment.

Geochemistry/Biogeochemistry & Subsurface Science



Unraveling
molecular-level
phenomena to
determine their
impact on
contaminant
migration and
transformation.

Science of Interfacial Phenomena



Developing & verifying predictive models for interfacial processes and advancing understanding of structure-function relationships in complex systems.

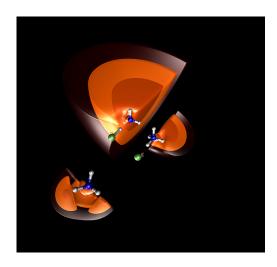
Environmental Molecular Sciences Laboratory: A National Scientific User Facility



Capabilities

EMSL houses an unparalleled collection of over 100 state-of-the-art imaging capabilities that are used to address scientific challenges:

- Molecular Science Computing
- Deposition and Microfabrication
- Kinetics and Reactions
- Mass Spectrometry
- Microscopy
- > NMR and EPR
- Spectroscopy and Diffraction
- Subsurface Flow and Transport



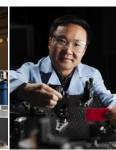












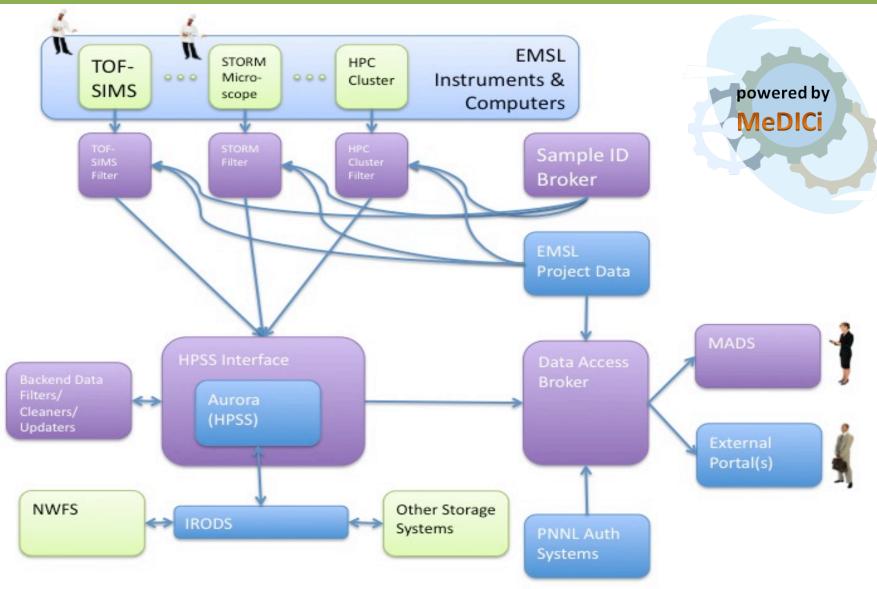






MyEMSL Base Architecture





What is the MeDICi Integration Framework?



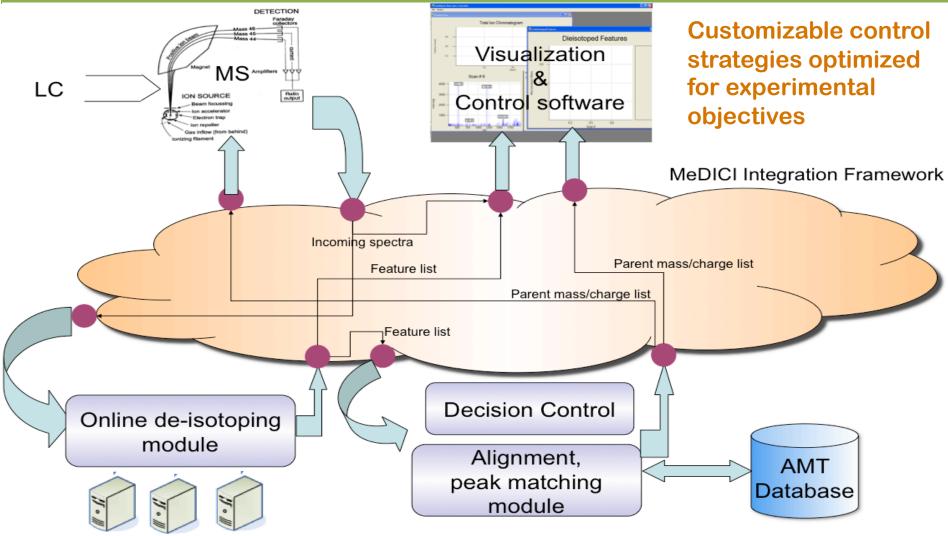
- Java-based integration technology
- Component-based API for creating analytical pipelines
 - Asynchronous component model for Java or non-Java (eg .exe, C/C++, R, Haskell, etc) codes (flexible)
 - Components can be distributed or executed in framework container (scalable through replication/partitioning)
 - Components communicate over a variety of protocols (e.g. JMS, Web Services, sockets, etc.) (configurable)
- Built on robust, industry-tested Java technologies
 - Service-Oriented Architecture (Mule open source ESB)
 - Java Messaging Service (e.g., ActiveMQ,)
- Hooks for provenance capture/workflow orchestration



Smart Instrument Control

High Performance Data Analysis Tools for Intelligent Mass Spectrometer Pipeline



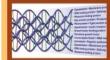


Next Generation Proteomics Pipeline



Data Genomes Metagenomes

ScalaBLAST



used by investigators to rapidly analyze large amounts of biological sequence data where detailed understanding of the molecular makeup of organisms or communities is crucial. ScalaBLAST provides scientists the information they need to map genes and pro-10000000000 and capabilities of the organisms. Poses bear character to the control of the control

Develop Sequences

Comparative sequence analyses are used to mine spectra, explore biological activity and identify species



Protein Families

Protein sequences are organized by similar ity, or "families," using advanced clustering algorithms which partition the metaproteom and metagenomes, and which allows for easier interpretation of the data. Grouping proteins by cellular function helps biologist address questions about what the system is doing. Similarities in sequences also indical an evolutionary relationship among microb ommunity members.



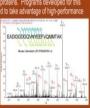
Peptide Identification

housands of mass spectra are generated for a single sample These spectra are automatically and routinely matched to sequences to identify proteins. Programs developed for this process are parallelized to take advantage of high-performance

computing for the purpose of speeding up the sample

MS/MS

MS



Visualization/Interaction

Interactive visual content analysis of real-time streams provides easily understandable visualizations of large datasets to systems biologists. This tool allows the user to visually navigate and explore data streams in manners not previously possible, making the search for relevant protein function much more dynamic and effective. Key

Data is incorporated into a visualization in real time.

the user's set

. Easy-to-understand visualization organizes data points and are

SMART INSTRUMENT CONTROL

Ion Mobility Spectrometry (IMS) Time of Flight

plexing schemes (such as Hadamard sforms) addresses the inherent under bling limitation of IMS-TOF measure-

The Smart Instrument Control

Technology facilitates intelligent

data gathering in tandem mass

spectrometry experiments by

analyzing acquired spectra on-



Field Programmable Gate Array (FPGA) Acquisition

IMS-TOF signal detection requires multidi mensional signal averaging of high speed streaming data. The current IMS-TOF streaming data. The current most 100 detection systems do not utilize the vendor's proprietary detection hardware due to the data intensive challenges presented in the architecture. DICI capability developments have enabled the MS-TOF team to form a collaboration with the vendor for an FPGA based solution to process the high speed streaming data. This research will result in IMS-TOF improvements in:

 Sensitivity · Data density

Mass

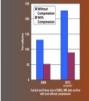
Tag DB

Resolution and mass measurement accuracy.

Data Compression

This technology is a compression algorithm that reduces data size while significantly increasing the speed for storing and extractin data. Key features include:

- . Data storage size reduction by a factor of 3 . Compression and storage of 240 MB of data in less than one second.
- Extract individual segments of uncompre data from a compressed file.



the-fly and performing pattern matching against a database of known signatures. Using this technology, researchers can reduce the amount of redundancy in their data collection procedures while also maximizing the information contained within the biological samples.



Feature Detection

Mass spectrometry features are extracted from raw instrument datasets and deisotoped - a process that involves the detection, deconvolution, and calculation of a monoisotopic mass rom clusters of peaks - representing different combinations of



Feature Identification

Mass spectrometry features are aligned from multiple datasets and compared to mass tag databases to make quantitative compansons of proteins from several samples at once. Addi-

studies involving seasonal fluctuations in protein turnover, and cellular pathways.

Mass Tag DB's



MeDICI

Middleware for Data Intensive Computing (MeDICi) is a software toolkit for creating scientific pipelines that can pull data from multiple instruments at a time and manipulate, format and organize the data for researchers to effectively view its information. Features are:

- Allowing scientists to visually create and modify a data processing pipeline to match their analysis needs.
- Supporting multiple programming languages, communication protocols, hardware platforms.
- Making it possible to capture and integrate high-throughput sensor data in real-time.

Biology Samples

Achieved to Date



- General infrastructure for automated data capture and annotation
- Intelligent Instrument Control linked to data storage and annotation through workflow framework
- Single Instrument automated analysis pipeline seamlessly linked to instrument control, data capture and annotation





Real Time Analysis and Integration – PNNL Chemical Imaging Initiative



Science needs to answer complex Questions

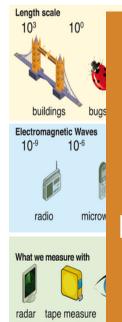


Investigative Methods

Result Representation

Methods

The many **Colours** of light



Many different investigative Methods

Many different result *Types* – Scale and Representation

Different methods deliver different *Insights* – more detail, new information

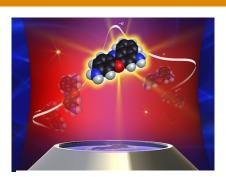
DIAMOND Light Source UK, 2010

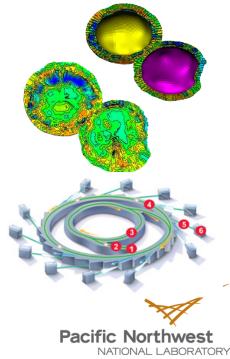




Real-World Manipulation on a Molecular Level

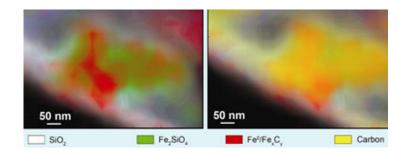
- Purpose: Deliver a suite of unique tools with nanometer scale resolution and element specificity that will allow researchers to go from model system observation to real-world manipulation on a molecular level.
- Approach: We will build signature, in-situ capabilities in
 - Light source based x-ray and VUV probes coupled with laboratory based imaging capabilities for 3D tomographic, structural, and element specific interrogation at the molecular level
 - Coupled optical, electron, ion, mass, and scanned probe microscopies to understand chemical and biological transformations and mechanisms
 - Integrative hardware and software applications for image reconstruction, feature extraction and information integration



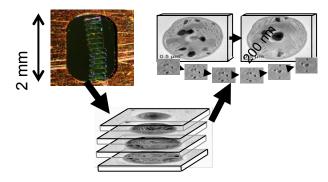


What will allow us to go from model system observation to real world manipulation of *in situ* interfaces on a molecular level?

- Direct visualization of chemical, material, and biological transformations are essential to achieve a confident level of control over complex systems
- Most of our ability to control matter today is through inference and interpretation of spectroscopic and structural data and modeling
- Direct Observation: With molecular scale imaging tools and concomitant data handling and analysis methods we can achieve the level of control enabling rational design and synthesis of new chemical, biological, and materials systems.



Chemical map of a FT catalyst – proto molecular movie



3D EM tomography of Cyanothece Cell



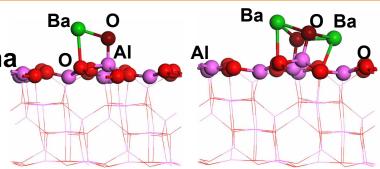
Atomic and Electronic Structure of Catalysts by Combination of In-situ and Ex-situ Imaging

Challenges:

- Oxide nanoclusters supported by gamma alumina
 - Physical structure and chemical state of elements
 - Electronic structure
 - Influence of these in catalytic properties

Approach:

- Class of TMO (e.g., WO₃, MO₃, V₂O₅) and TM (e.g., Pt, Cu) supported by g-Al₂O₃ substrate are important
- Combination of in-situ and ex-situ aberration corrected TEM, STEM-HAADF imaging, EDS and EELS spectroscopy
- Integration of high energy resolution EELS measurements with light source XAS
- Integration of experimental data with DFT calculations



BaO monomer and dimer on a dehydroxylated γ -Al₂O₃(100) surface

Kwak et al, Journal of Catalysis 261, 17-22, 2009



Proudly Operated by Battelle Since 1965

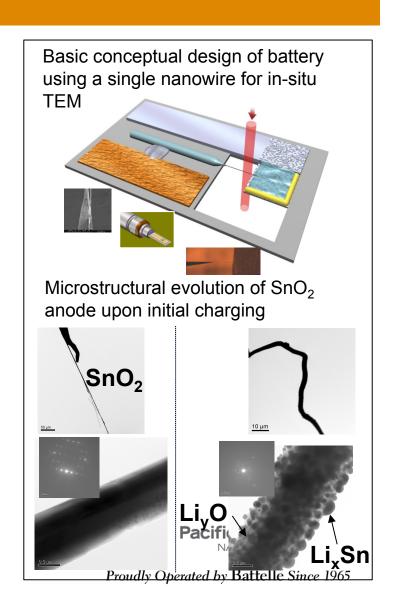
Understanding Chemical and Structural Changes in Battery Materials

Challenges:

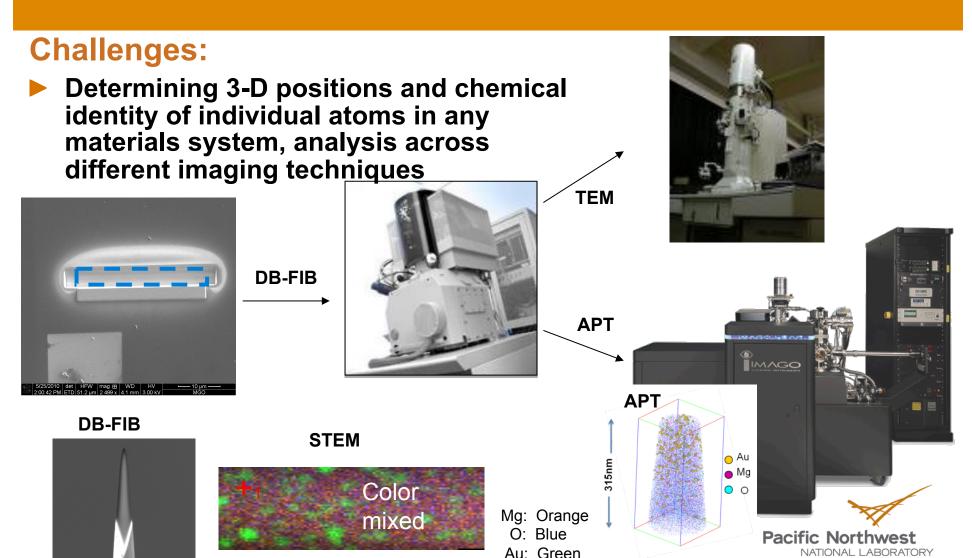
- Discovering new materials with high capacity
 - Less volume expansion
 - Enhance charging and discharging rate
 - Less irreversible microstructure formations

Approach:

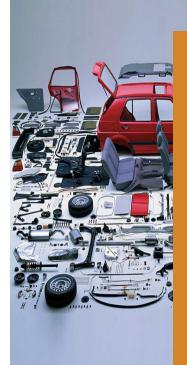
- Combination of nanomaterials and in-situ structural and electrochemical measurements
- Combination of TEM, STEM and APT
- Integration of high energy resolution EELS measurements with light source XAS techniques chemical state charge transfer information
- Integration of experimental data with calculations (DFT and MD)



Integrated Sample Preparation and Analysis Platform for Comprehensive 3-D Chemical Imaging



Effectively relating Information and Knowledge



We need a map and plan that describes how things fit together - in specific areas and in the whole

Different people have different perspectives and experiences, expecting information represented in their context



Context sensitive, flexible Framework



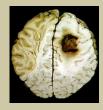
- Describe
- •Map
- Classify



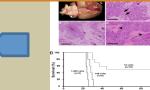
We need a flexible framework that allows scientists to integrate, compare and contrast results from different investigative methods – and presents the results in a context that they are familiar with

Characterize Different Components

Framework Relating Comparing Combining **Synthesizing**















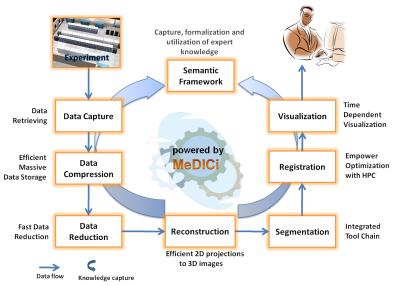
View Points -Shared Paci Knowledge

Different

erated by Battelle Since 1965

Develop synergistic integration of multiple imaging, characterization, and simulation techniques

- Phase 1: In collaboration with scientists establish basic data analysis and handling framework
- Phase 2: Develop integrative analysis methods across different chemical imaging technologies with support of the wider community
- Phase 3: Evolve data handling and analysis methods to meet the real time and high data volume requirements of the new chemical imaging capabilities

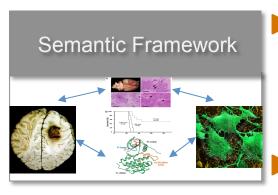




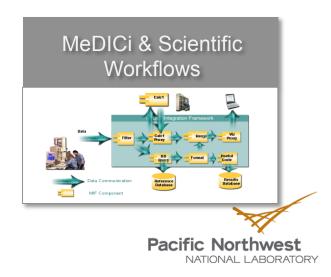
Key Challenges

Challenge	Today	Tomorrow	Technologies
High Data Rate	5PB / Year LHC	3.5 PB / Day XFEL	Storage / Movement/ Analysis
Real Time Analysis	After Experiment -18 hours for 3D TEM reconstruction	Experimental Steering through real time analysis during experiment	Real Time, Parallel, Data Intensive Computing HW + SW
Integration across different Imaging Technologies	One off integration of 2 techniques	100's of combinations possible	Conceptual Relation, real time integration
Integrating across scales	On related scales only	Nano to Macro	Conceptual Relation
Geographical distribution of experimental sources to be integrated	None	10's – flexibly combined	Distributed Computing Paradigm

Proposed Core Framework Development (1)

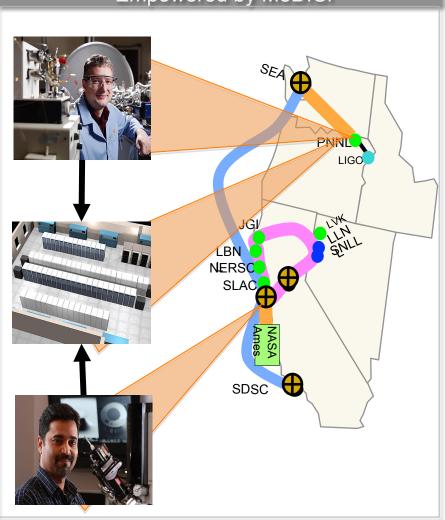


- Formal characterizations of the methods, instruments, samples, analysis processes and associated data products.
- Formalized topology of the methods, their contribution and constraints.
- Flexible creation of data intensive workflows.
- Managing complex and intensive data exchange as well as rapid integration of data sets spanning different spatial and temporal scales



Proposed Core Framework Development

Data Capture and Distributed Computing Empowered by MeDICi



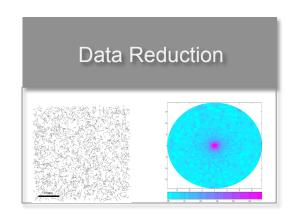
- Leveraging MyEMSL framework providing workflow, data capture, metadata capture, a central data repository, and tools for data discovery.
- High volume data transfers.
- Taking analysis to the data via distributed computing.

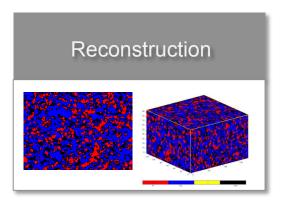


Proposed Example Components (1)



- Appropriate lossless and lossy compression algorithms
- High compression ratio with low computational overhead
- Reduce noise and smooth data
- Reconstructions will contain the most significant information, are featureaccentuated

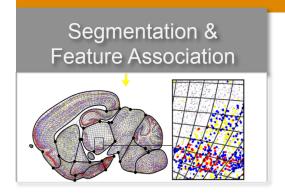




- Accurate re-construction of high volume data in real time
- Combine correlation functions with parallelized filtered back projection

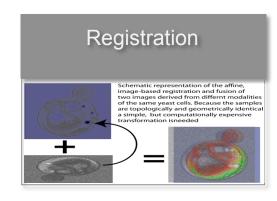
ЗY

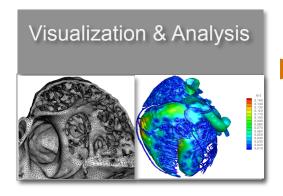
Proposed Example Components(2)



Scalable differential operators and primitives that can be combined at run time for application-specific chemical signature and feature recognition

Enabling accurate localized comparisons between experimental datasets from different chemical imaging techniques at high resolutions





Real time, remote, in-situ, high data volume 3+4D visualization

NATIONAL LABORATORY

Initial Steps

- Identification of a small number of key scientific workflows
- Determine analysis and integration steps together with scientists
- Implement initial end to end workflows to exercise all framework components
- Move on to other workflows test generalization, encourage community involvement



Questions?